

## A BLOCKING STRATEGY

C. Vithayasai and D. S. Robson

Computing and Applied Statistics Directorate,  
Environment Canada, Ottawa  
Biometrics Unit, Cornell University, Ithaca, New York

### ABSTRACT

We suppose that prior to partitioning a set of experimental units into experimental blocks of size  $t$ , each unit is classified with respect to  $k$  different binary variables or attributes. The  $2^k$  cells of this classification are then used as building blocks to form experimental blocks of  $t$  homogeneous experimental units. In our attempt to achieve within-block homogeneity we first order the  $2^k$  cells in a sequence so that any two adjacent cells differ with respect to exactly one attribute. The contents of each cell are then partitioned into blocks and, where necessary, units from adjacent cells are combined to form blocks. In the case of standardized quantitative variables which have been dichotomized at zero, a within-cell partition is based upon an ordering of the experimental units with respect to that particular quantitative variable which changes its sign in the next cell of the sequence.

### INTRODUCTION

Blocking is an experimental design technique universally employed as a means of reducing the magnitude of experimental error.

Experimental error variance among a set of experimental units is understood to measure the variance in response which would obtain if all these experimental units were treated alike. In an experimental comparison among  $t$  treatments, each applied to a different set of experimental units, the difference in response between any two treatments is confounded with the inherent differences in the two sets of experimental units.

These inherent differences between the two sets treated differently can often be reduced in a deliberate manner by the process of blocking. Prior to the assignment of treatments, the available experimental units may be partitioned into blocks in such a manner that units within each block are as homogeneous as possible. If different treatments are then assigned to the units within a block the subsequent differences in response should more accurately reflect the treatment effects. In the case of a randomized complete block design with  $r$  replications of  $t$  treatments, for example, the  $rt$  experimental units are first partitioned into  $r$  blocks of  $t$  units each and then the  $t$  different treatments are randomly assigned to the  $t$  units in each block.

The degree of success of such a blocking procedure is measured by the reduction achieved in the magnitude of experimental error. Without blocking, the experimental error variance is simply the total variance in response among the  $rt$  experimental units were they all to receive the same treatment; in the blocked experiment that portion of the total variance attributable to differences between blocks is eliminated from the experimental error variance, which then becomes the pooled within-block variance.

The key to success, of course, lies in the experimenter's ability to anticipate which units would respond alike if treated alike. In the case of field plot experiments this problem is relatively simple; contiguous plots have many environmental features in common and hence blocks are formed from geographically contiguous plots. In animal experiments, litter mates have similar genetic and environmental backgrounds and hence may be expected

to respond similarly to treatment. Since the theory of experimental design and analysis has its origins in agricultural experimentation where the most effective blocking strategy is frequently self-evident, little attention was or has since been given to the fine structure of objective blocking strategies. Outside of such obvious cases, however, the blocking problem can become both perplexing and, as in the case of clinical trials, critically important. While literature on this topic does exist (see, for example, Rubin [1973]) a paucity of methodology also exists for a procedure which is so basic to the design of experiments.

Our approach to this problem is strictly heuristic and is confined to a situation in which each of the  $rt$  experimental units has been measured with respect to  $k$  different quantitative variables which are presumed to have been selected for their close relationships with one or more of the intended response variables. Our objective then is to develop an algorithm for partitioning the  $rt$  experimental units into  $r$  blocks of size  $t$  so that the units within a block are as much alike as possible with respect to all  $k$  variables. The method we propose has no known optimality properties but has been demonstrated to operate with reasonable effectiveness in some computer simulations.

#### BLOCKING STRATEGY

Before proceeding with the case of quantitative variables  $\underline{x} = (x_1, \dots, x_k)$  we consider the case of attributes or binary variables where  $x_j = 0$  or  $1$  for  $j = 1, \dots, k$ . The  $rt$  units are then distributed over  $2^k$  cells which may be regarded as building blocks for our purpose of constructing experimental blocks, each building block consisting of a set of identical experimental units (identical with respect to the binary vector variable  $\underline{x}$ ). A cell containing more than  $t$  units will then have to be partitioned, while if a cell contains fewer than  $t$  units then additional units will have to be drawn from a neighboring cell or cells in order to form a complete block of  $t$  units. In the latter case a nearest neighboring

cell is one which differs from the cell in question in only one dimension; i.e., if the cell in question bears the binary label  $(\delta_1, \dots, \delta_k)$  where each  $\delta$  is either 0 or 1, then the  $k$  nearest neighboring cells are those bearing the labels  $(\delta_1, \dots, 1-\delta_j, \dots, \delta_k)$  for  $j = 1, \dots, k$ .

Block construction in the binary case could then proceed in this manner starting, say, with the cell  $(0, 0, \dots, 0)$  and forming as many blocks as possible within this cell then carrying the remainder along to combine with units from a neighboring cell. The only serious problem here is the specification of a path through the array of  $2^k$  cells, and to this end we point out that the possible paths are characterized by the sequences:

$k = 1:$  I, a  
 $k = 2:$  I, a, ab, b  
 $k = 3:$  I, a, ab, b, bc, abc, ac, c  
 $k = 4:$  I, a, ab, b, bc, abc, ac, c, cd, acd, abcd, bcd,  
 bd, abd, ad, d

which continue in an obvious manner. Thus, for  $k = 3$ , one possible path starting from cell (000) is:

$(000) \rightarrow (100) \rightarrow (110) \rightarrow (010) \rightarrow (011) \rightarrow (111) \rightarrow (101) \rightarrow (001)$

while another is:

$(000) \rightarrow (010) \rightarrow (110) \rightarrow (100) \rightarrow (101) \rightarrow (111) \rightarrow (011) \rightarrow (001)$ .

The second path, however, can be obtained from the first merely by interchanging the variables  $x_1$  and  $x_2$ . In this sense all nearest neighbor sequences are equivalent to the first path, being derivable from the first by a permutation of the three variables, and we shall therefore refer to the first as the generic ordering. The generic order number  $g$  of a cell with the binary label  $\underline{\delta} = (\delta_1, \dots, \delta_k)$  is given by the formula

$$g(\underline{\delta}) = 2^{k-1} + \frac{1}{2} \left[ 1 - \sum_{j=1}^k 2^{j-1} (-1)^{\delta_j + \delta_{j+1} + \dots + \delta_k} \right]$$

and the  $i$ 'th vector  $\underline{\delta}_i = (\delta_{i1}, \dots, \delta_{ik})$  in the generic sequence  $\underline{\delta}_1, \underline{\delta}_2, \dots, \underline{\delta}_{2^k}$  therefore satisfies the equation  $g(\underline{\delta}_i) = i$ . Note that the generic ordering is circular in the sense that the first and last cells are nearest neighbors.

Continuous variables  $x_j$  can always be transformed into binary variables by letting  $\delta_j = 1$  or  $0$  according as  $x_j > c_j$  or  $x_j \leq c_j$  for any specified  $c_j$  as, for example, when  $c_j$  is the mean or median value of  $x_j$  among the  $t$  experimental units. The members of any one of the resulting  $2^k$  cells would no longer be identical with respect to the vector variable  $\underline{x}$ , however, and if the number of units in the cell exceeded  $t$  then the values of  $\underline{x}$  would have to be examined to identify a subset of  $t$  units which are in some sense homogeneous with respect to  $\underline{x}$ . Looking ahead we note that once this cell has been blocked it will be necessary to move on to a specified nearest neighboring cell, carrying along any remaining units from the first cell. The major difference between the two cells is that for some single  $j$ ,  $x_j$  lies on either side of  $c_j$ ; thus, the remaining units which we should be carrying along are those with  $x_j$  closest to  $c_j$ . This line of reasoning thus suggests that blocking within the first cell should be based on the rank order of the units with respect to the quantitative variable  $x_j$ .

Figure 1 illustrates an application of this algorithm in the two-dimensional case and Figure 2 presents a flow chart for computer programming the  $k$  dimensional case using the following definitions of variables:

$B$  = block identification number for the block currently under construction,  $B = 1, 2, \dots, r$  (initial value of  $B = 0$ )

$R$  = number of experimental units from preceding cells which have already been assigned to the current block,  
 $0 \leq R < t$  (initial value of  $R = 0$ )

$i$  = generic order number of the cell currently under consideration,  $i = 1, 2, 3, \dots, 2^k$  (initial value of  $i = 1$ )

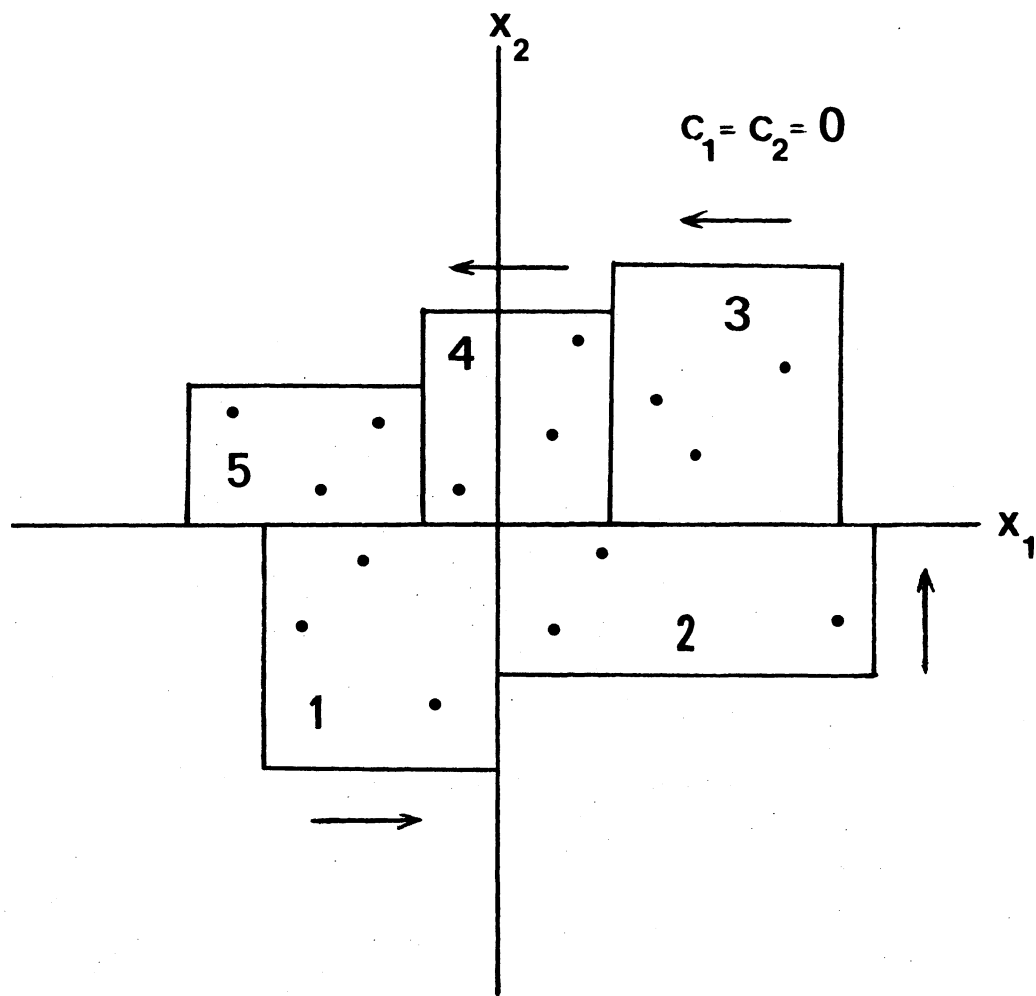


FIG. 1

Schematic implementation of the blocking algorithm  
in the two-dimensional case with  $r=5$  and  $t=3$ .

Flow chart for the blocking algorithm.

$N_i$  = number of experimental units initially contained in the  
i'th cell

$$M = N_i + R$$

IP = integer part of  $(M/t)$   
= number of complete blocks to be formed from M units

$(\delta_{i1}, \dots, \delta_{ik})$  = binary label of cell i

J = identification number of the variable on which the ex-  
perimental units in the current cell are to be ordered

S = the direction of ordering, either + or - :

if S = + then the units are ordered from the smallest  
 $x_j$  to the largest  $x_j$

if S = - then the units are ordered from the largest  
 $x_j$  to the smallest  $x_j$

The flow chart applies after the cutting points  $\underline{c} = (c_1, \dots, c_k)$  have been specified and the data vector  $\underline{x} = (x_1, \dots, x_k)$  of each experimental unit has been classified with respect to  $\underline{c}$  and stored in a cell bearing both the binary label  $\underline{\delta} = (\delta_1, \dots, \delta_k)$  and the generic order number  $g(\underline{\delta})$ , where  $\delta_j = 0$  or 1 according as  $x_j \leq c_j$  or  $x_j > c_j$ . This preliminary processing of the data thus consists of either reading in a specified  $\underline{c}$  or computing  $\underline{c}$  as the mean or median vector  $\underline{x}$ , then computing  $\underline{\delta}(\underline{x})$  and  $g(\underline{\delta})$ , and storing and counting data in the  $2^k$  cells. The flow chart in Figure 2 is then entered with the initial values  $B = 0$ ,  $R = 0$  and  $i = 1$ .

A numerical example is shown in Table I for the three-dimensional case with  $r = 5$  and  $t = 6$ . The  $rt = 30$  ordered triplets of data represent 90 independent, random normal deviates from a population with mean zero and variance  $\sigma^2 = 4$ . A reduction from  $\sigma^2 = 4$  should therefore be expected as a result of blocking, and in this instance the average within-block variance was 2.57. The cutting points were here specified to be  $\underline{c} = (0, 0, 0)$ , and in Table I the data are displayed in the resulting generically ordered cells to permit application of the flow chart in Figure 2.



TABLE I

A Numerical Example Illustrating Application of the Blocking Algorithm for k=3, r=5 and t=6

000			100			110			010			011			111			101			001		
x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>
- .7	- .9	-1.5	1.0	-2.4	-2.5	1.2	2.0	-4.4	-1.6	.02	-2.7	- .6	1.7	.8	1.9	.8	.4	1.0	-1.9	1.5	-1.8	-1.9	.5
-3.2	- .9	-1.3				1.9	.5	-1.4	-.07	4.5	-.9	-2.4	3.2	.5				2.9	-4.7	5.0	-1.2	-.7	.2
-2.0	- .9	-1.7				1.6	.9	-2.0	-4.1	.09	-.6	-.07	.3	1.3				5.5	-3.1	.1	-.8	-.3	1.0
-4.0	-1.7	-.7				Block 2			-3.9	2.7	-.8	-.9	1.3	.3				1.6	-2.3	2.0			
-2.0	-2.7	-1.4							Block 3			-.4	.2	3.1				.9	-2.2	2.8	Block 5		
-1.3	-3.9	-.4										-1.2	.8	3.2									
-1.0	-5.6	-3.9													Block 4								
Block 1																							

Block															
1			2			3			4			5			
x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	
-3.2	- .9	-1.3	- .7	- .9	-1.5	-1.6	.02	-2.7	- .07	.3	1.3	2.9	-4.7	5.0	
-2.0	- .9	-1.7	1.0	-2.4	-2.5	-4.1	.09	- .6	- .4	.2	3.1	5.5	-3.1	.1	
-4.0	-1.7	- .7	1.2	2.0	-4.4	-3.9	2.7	- .8	-1.2	.8	3.2	1.6	-2.3	2.0	
-2.0	-2.7	-1.4	1.9	.5	-1.4	- .6	1.7	.8	1.9	.8	.4	-1.8	-1.9	.5	
-1.3	-3.9	- .4	1.6	.9	-2.0	-2.4	3.2	.5	1.0	-1.9	1.5	-1.2	- .7	.2	
-1.0	-5.6	-3.9	- .07	4.5	- .9	- .9	1.3	.3	.9	-2.2	2.8	- .8	- .3	1.0	
Block mean	-2.3	-2.6	-1.6	.8	.8	-2.1	-2.3	1.5	- .4	.4	- .3	2.1	1.0	-2.2	1.5
Variance	1.31	3.46	1.53	.98	5.67	1.55	2.23	1.78	1.65	1.24	1.84	1.32	8.03	2.60	3.48
									x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	Average			
Average Within-Block Variance									2.76	3.07	1.91	2.57			

## DISCUSSION

When each experimental unit has been measured with respect to several concomitant variables as well as a response variable, a standard statistical method for eliminating the nuisance effect of variation in the concomitant variables is the analysis of covariance. Availability of this statistical method of accounting for and removing the effects of measured differences between experimental units may in part explain the paucity of blocking methodology for quantitative variables.

When variation in such a potent, composite factor as geographic location is available to the plant experimenter, for example, then a strong case can certainly be made for blocking on this factor and using covariance analysis to remove the effect of other measured, concomitant variables. The target situation for our blocking algorithm, however, is one in which the only relevant information available to the blocking strategist is the measurement  $\tilde{x}$  on each experimental unit.

In such circumstances total reliance upon a data analytic method of adjusting treatment differences for differences in concomitant variables should be avoided whenever possible on the grounds that such methods are parametric and rely upon assumptions which are often untestable. Both the covariance analysis for removing the effect of  $\tilde{x}$  and the conventional data analysis of an experiment blocked on  $\tilde{x}$  are based upon the assumption of additivity (no interaction between the treatment factor and the concomitant factors measured by  $\tilde{x}$ ), but covariance analysis further requires specification of the functional form of the regression of the response variable on  $\tilde{x}$ . The two techniques may, of course, be applied in series; in a randomized complete block design, however, blocking on  $\tilde{x}$  should minimize the treatment differences with respect to  $\tilde{x}$  and hence minimize the covariance adjustments of response differences for differences in  $\tilde{x}$ .

The statistical properties of this proposed blocking scheme have not yet been examined in any detail. Some rudimentary results

are available for the case where the components of  $\tilde{x}$  are independent normal deviates and the cutting points  $\tilde{c}$  are taken to be zero, a situation which is approximated when an ortho-normal transformation is made prior to blocking. In a block containing experimental units which were ordered with respect to one variable in the construction process, the within-block variance of each of the other variables is reduced to a fraction  $1 - \frac{2}{\pi} = .3633$  of total variance. This 64% reduction should approximately obtain for all components of  $\tilde{x}$  in the case  $r = 2^k$  where each of the  $2^k$  cells is then expected to produce one complete block, within which each component of  $\tilde{x}$  is homogeneous in sign and hence has variance  $1 - \frac{2}{\pi}$ . The reduction will be greater when  $r > 2^k$  and less when  $r < 2^k$ , but the exact relationship is unknown. Our numerical example in Table I with  $r = 5 < 2^k = 8$  produced an average reduction of 46.6% when the observed within-block variance was compared to the observed total variance of each variable.

Similar statements can be made with respect to a normally distributed response variable; if the ortho-normal variable  $\tilde{x}$  accounts for a fraction  $R^2$  of the total variability in response then when  $r = 2^k$  the experimental error should be reduced approximately  $64R^2\%$  by this blocking procedure. In canonical form the regression of a response variable  $y$  on the ortho-normal variable  $\tilde{x}$  is

$$y = \rho_{y\tilde{x}_1} \tilde{x}_1 + \dots + \rho_{y\tilde{x}_k} \tilde{x}_k + \epsilon$$

with conditional variance  $\text{Var}(y|\tilde{x}) = \sigma_\epsilon^2$  given by

$$\sigma_\epsilon^2 = 1 - \rho_{y\tilde{x}_1}^2 - \dots - \rho_{y\tilde{x}_k}^2 \equiv 1 - R^2.$$

The within-block variance of  $y$ , say  $V_w(y)$ , is then given by

$$V_w(y) = 1 - R^2 + \rho_{y\tilde{x}_1}^2 V_w(\tilde{x}_1) + \dots + \rho_{y\tilde{x}_k}^2 V_w(\tilde{x}_k)$$

so if  $V_w(\tilde{x}_j)$  is simply the conditional variance

$$V_w(\tilde{x}_j) = \text{Var}(\tilde{x}_j | \text{sign}[\tilde{x}_j]) = 1 - \frac{2}{\pi}$$

for all  $j$  then  $V_w(y) = 1 - \frac{2}{\pi} R^2$ .

Other aspects of this procedure which require investigation are the effects of permuting  $\tilde{x}$  when the components of  $\tilde{x}$  are not of equal importance, and possible extensions of the procedure such as the use of  $p^k$  classes or more generally  $p_1^{k_1} p_2^{k_2} \dots p_m^{k_m}$  classes when the  $k = k_1 + \dots + k_m$  variables are not of equal importance. Rearranging the components of  $\tilde{x}$  would seemingly have a negligible effect on the resulting within-block variance of each component when  $r \geq 2^k$ , but for smaller values of  $r$  the variance reduction for each component clearly depends on its position in the  $x$ -vector. The extension to a factorial array of classes with ordered levels of each factor is readily visualized in three dimensions, but devising a pathway algorithm for the general case is a nontrivial algebraic problem, and a statistically rational criterion for ordering the components of  $\tilde{x}$  in this general case appears to be an imponderable problem.

Still another statistical problem is the development of a criterion for culling blocks or experimental units when either is available in excess of requirements. One consideration in this regard is the objective of achieving not only reduced within-block variance but also homogeneity of within-block variance for the response variables. Variance homogeneity in the concomitant variables does not imply variance homogeneity in the response variables unless the regression functions are linear, but aberrant blocks such as block 5 in our numerical example could well be culled if a surplus exists.

#### BIBLIOGRAPHY

Rubin, D. B. (1973). Matching to remove bias in observational studies. Biometrics 29, 159-183.